DOCUMENT RESUME

ED 347 173 TM 018 539

AUTHOR Dolenz, Beverly
TITLE Factors That Attenuate the Correlation Coefficient and Its Analogs.
PUB DATE 21 Jan 92
NOTE 20p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Houston, TX, January 31-February 2, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; *Correlation; Heuristics; *Relationship; Reliability; *Statistical Distributions
IDENTIFIERS *Attenuation (Paradox); Linear Relationships; Scattergrams; Variables

ABSTRACT

The correlation coefficient is an integral part of many other statistical techniques (analysis of variance, t-tests, etc.), since all analytic methods are actually correlational (G. V. Glass and K. D. Hopkins, 1984). The correlation coefficient is a statistical summary that represents the degree and direction of relationship between two variables. It is important to understand the various factors that affect this coefficient. This paper explains the assumptions underlying the use of the bivariate correlation coefficient, using small heuristic data sets and graphs to make the discussion concrete and accessible to the reader. Attenuation influences on the coefficient of correlation include: (1) departures from linearity; (2) departures from both variables being similarly distributed; (3) using instruments with lower reliability; and (4) using data in which either variable has a restricted range. Six tables and six scatterplots present study data. There is a nine-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

bruce.wp1 1/21/92

ED347173

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BEVERLY DOLENZ

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Factors That Attenuate the Correlation Coefficient and its Analogs

Beverly Dolenz
Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX, January 31, 1992.

TM018539

BEST COPY AVAILABLE

ABSTRACT

The correlation coefficient is an integral part of many other statistical techniques (ANOVA, *t*-tests, etc.), since all analytic methods are actually correlational (Glass & Hopkins, 1984). Given the importance of *r*, it is important to understand the various factors that affect *r*. The present paper explains the assumptions underlying the use of the bivariate correlation coefficient, using small heuristic data sets and graphs to make the discussion concrete and accessible to the reader.

The coefficient of correlation, r, is a statistical summary which represents the degree and direction of relationship between two variables (Glass & Hopkins, 1984). It is a measure of the strength of *linear* association between the two variables (Busby & Thompson, 1990) and is a developmental basis for many other statistical techniques (Glass & Hopkins, 1984).

Indeed, all parametric tests can be treated as special cases of canonical correlation analysis, which is the most general parametric procedure for investigating the relationships between two sets of variables (Baggaley, 1981, Fornell, 1978; Knapp, 1978; Thompson, 1991). Parametric statistics (t-tests, ANOVA, ANCOVA, regression, etc.) are all different indices of relationship, with the commonality that assumptions about the distributions of variables may be required (Nunnally, 1967).

Given the importance of r as a methodological concept that links all parametric statistics, it is important for researchers to understand the various factors which attenuate r. In this paper, assumptions associated with r are reviewed, factors which attenuate r are presented, and a small data set is statistically manipulated to further demonstrate the influence of particular factors on r.

Assumptions

Certain assumptions are associated with r, just as assumptions are associated with virtually all statistics. Nunnally (1967) outlines these assumptions as follows: 1) there must be a linear relationship between variables, i.e., a straight line must do a good job of describing the trend in a bivariate scattergram of the

two variables regardless of how much points may scatter above and below this line; 2) both variables must be distributed similarly, i.e., must have the same shape; and 3) the relationship between variables must be homoscedastic, i.e., the spread about the best-fitting straight line in a scattergram of the variables must be approximately the same at all levels of the two variables. This last assumption is primarily important if one is testing the statistical significance of r.

According to Nunnally (1967), unless one of the assumptions is seriously violated, inferential statistical tests of the statistical significance of r will yield reasonably accurate conclusions. Nunnally (1967) sights as an example of a "serious" violation of an assumption the correlation of a normally distributed variable with scores from a distinctly J-shaped distribution. If there is some evidence of departure from the assumptions, one plausible procedure is to use a higher level of "significance" than ordinarily would be required (Nunnally, 1967), but the effect size (r^2) estimate will still be attenuated.

Factors That Influence r

The interpretation of the correlation coefficient is influenced by many factors. Attenuation influences on r include departures from linearity, departures from both variables being similarly distributed, using instruments with lower reliability, and using data in which either variable has a restricted range (e.g., only gifted students provided scores on an IQ variable being correlated with a second variable).

Linearity

A coefficient of correlation is a measure of the linear relationship between two variables. The relationship between two variables is linear if a "best-fitting" curved line does no better at fitting the data points in the scattergram than does the "best-fitting" straight line, called the regression line (Glass & Hopkins, 1984). To evaluate the nature of the relationship between two variables, a researchers can study a scatter diagram and "eyeball" the plot to surmise whether or not the relationship between X and Y is basically linear. If X and Y have any degree of curvilinear relationship, the value of r will underestimate the true degree of relationship between the two variables (Glass & Hopkins, 1984).

A scatterplot enables a researcher to surmise whether or not a computed r accurately summarizes the relationship between the two variables. The value of r is solely a measure of the degree of linear relationship between X and Y, except as r approaches its limits (+1 or -1), in which case the researcher can then be confident that no curvilinearity is present. If the actual relationship is curvilinear, however, the value of r may be erroneously close to 0, even though considerable curvilinear relationship between the variables is present. This illustrates the importance of always examining scatterplots whenever the relationship between variables is studied using r.

Table 1 presents hypothetical data that are curvilinearly related, and descriptive statistics for the data are presented in

Table 2. The hypothetical example involves a relationship between anxiety and test scores. The resulting r for these data is .0, suggesting no relationship between the variables. But a quick look at the Figure 1 scattergram, which plots the relationship between the two variables, demonstrates that a relationship does exist between the variables, but not a linear one.

INSERT TABLES 1 AND 2 AND FIGURE 1 ABOUT HERE

Departures from Similar Distribution Shapes

Departures from similar distribution shapes can also result in conservative underestimates of relationship (Nunnally, 1967). Thus, unless two variables have exactly the same distribution, it is simply not possible to obtain a perfect correlation (+1 or -1) between the two variables.

The maximum value of r is 1.0. However, r can equal 1.0 only when the marginal distributions of X and Y have precisely the same shape (Glass & Hopkins, 1984). If X is normally distributed and Y is negatively skewed, the maximum value of r is less than 1. The more the shapes differ, the more the value for r will be restricted.

Table 3 presents hypothetical data involving SAT math scores and scores from a highly correlated math test. The data are manipulated here to demonstrate the effect that the shapes of distributions have upon r. Notice that the "SATmath" (X) and the "Mathtest" (Y) variables have similar shapes, as evidenced by their having the same coefficients of skewness (.00), as noted in Table

4, and are perfectly correlated. Figure 2 graphically demonstrates this.

INSERT TABLES 3 AND 4 AND FIGURE 2 ABOUT HERE

"SATmath" (X) remains normally distributed as it is correlated with "Yskewed", the math test scores which are manipulated to form a negatively skewed distribution. Notice in Table 4 the marked decrease in r when correlating variables with dissimilar shapes. The second scattergram in Figure 2 graphically demonstrates this.

The degree of restriction on the correlation coefficient depends on how high the correlation would be if the distributions had the same shapes (Nunnally, 1967). Small increases in r will be anticipated when linear transformations in the data are conducted to make the distributions more similar in shape (Glass & Hopkins, 1984).

Imperfect Measurement Reliability

Another factor which places a limit on the value of r involves the measurements used in a study (Glass & Hopkins, 1984). Measurement error lends to attenuation of r (Busby & Thompson, 1990). Because the reliability coefficients of the two variables being correlated establishes a ceiling for r for a given data set, and because it is incorrect to describe tests as being reliable (Thompson, 1992), it is important to always assess the reliability of the scores in hand on both variables being correlated. As Busby and Thompson (1990) explain, r can never exceed the square root of the product of the two reliability coefficients for the scores in

hand.

Restricted Range

One of the most noteworthy influences on r is the variance in the sample (Glass & Hopkins, 1984). Nunnally (1967) asserts that if a broader range of subjects is studied, the correlation will increase; if a narrower range of subjects is studied, the correlation will decrease.

The change in variance must be a real change, however, arising from differences in sampling methods, such as initially selecting the top 50% of scores on the SAT math exam to correlate with math scores on a separate test. Artificial changes in the variance, as would occur if all the scores on the dependent variable were multiplied by five, does not ultimately change z-score versions of the scores, and consequently correlation coefficients between the scores will remain the same.

Table 3 presents the "Ytimes5" data and Figure 3 graphically demonstrates that artificially changing the variance in a sample does not affect r. Notice that the "Mathtest" (Y) scaled scores have been altered into "Ytimes5" in both Table 3 and Figure 3, but r is unaffected.

INSERT FIGURE 3 ABOUT HERE

The prevalent concern in correlational analysis is with sampling methods that restrict the variance, referred to as restriction in range (Nunnally, 1967), which consequently drops the magnitude of r. Subject pools that are homogeneous underestimate

the magnitude of the relationship between the variables and represent a restricted sample. The restriction may be either direct or indirect (Nunnally, 1967).

A direct restriction occurs when sampling procedures are biased with respect to one of the two variables. Nunnally (1967) uses as an example an aptitude test that is being validated for the selection of college freshman; the test is administered to all students applying for admission to a particular college. Only the top 50% of the students on the aptitude test are selected for admission. Later, scores on the aptitude test of those admitted to college are correlated with grade-point-averages. There has been a direct restriction of the range on the aptitude test. The variance of scores on the test would have been much larger had all the applicants been admitted to college. Consequently, the validity coefficients obtained in the restricted sample will be spuriously low.

Table 5 presents an analogous example involving full and restricted ranges for hypothetical data involving SAT math scores and test scores from a math test to illustrate the results of direct restriction of range on r. In calculating r, the data are directly restricted utilizing the top 50% of "SATmath" scores. Table 6 presents the descriptive statistics and r for Table 5 data. Notice the change in the shape of the distribution, as evidenced by skewness, when the variance of the data is restricted. Also notice the decrease in r for the restricted sample.

INSERT TABLES 5 AND 6 ABOUT HERE

The first scattergram in Figure 4 plots the relationship between variables using the unrestricted data from Table 5. The horizontal and vertical reference lines represent the points of restriction. The upper right hand block of data points in the top scattergram in Figure 4 represents what is graphically presented in the second scattergram in Figure 4, i.e., the scattergram for the data when the range is restricted. Notice the information which is lost when data are restricted.

INSERT FIGURE 4 ABOUT HERE

Indirect restriction of range occurs when the restriction is on a third or extraneous variable that correlates with one of the two variables actually of primary interest in the study (Nunnally, 1967). If an extraneous variable is highly correlated with either of the two variables being investigated, and if the range of the extraneous variable is restricted in the study, then the *r* between the two variables of interest will still be attenuated.

Correction for Attenuation

Attenuation, the reduction of *r* as a result of sampling and/or methodological errors, can be corrected using statistical estimation formulas, but some controversy exists about when the correction for attenuation should be applied (Busby & Thompson, 1990). Correction for attenuation is not as useful in applied research, because the corrections themselves are fallible,

therefore, the use of correction formulas can exacerbate estimation or prediction problems (Hardyck & Petrinovich, 1969).

Glass and Hopkins (1984) present a correction for restricted range for estimating the unrestricted group correlation if the true standard deviation of the unrestricted variable is known. Glass and Hopkins (1984) stress that the purpose of this estimation is basically in illuminating the substantial consequence of restricted (or exaggerated) variability on the value of the observed correlation coefficient. This correction is most appropriate for large samples where the standard deviation and correlation coefficient are very accurate estimates of the population parameters (Glass & Hopkins, 1984).

Summary

The correlation coefficient is an integral part of many other statistical techniques (Glass & Hopkins, 1984). Given the importance of *r*, it is important to understand the various factors that affect *r*. The present paper has explained the assumptions underlying the use of the bivariate correlation coefficient, using small heuristic data sets and graphs to make the discussion concrete and accessible to the reader.

References

Baggaley, A. R. (1981). Multivariate analysis: An introduction for consumers of behavioral research. Evaluation Review, 5, 123-131.

Busby, D., & Thompson, B. (1990, January). Factors attenuating Pearson's r: A review of basics and some corrections. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

Fornell, C. (1978). Three approaches to canonical analysis. Journal of the Market Research Society, 20, 166-181.

Glass, G.V, & Hopkins, K.D. (1984). Statistical methods in education and psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hardyck, C.D., & Petrinoviach, L.F. (1969). Introduction to statistics for the behavioral sciences. Philadelphia: W.B. Sanders.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Nunnally, J.C. (1967). Psychometric theory. New York: McGraw-Hill Book Company.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24(2), 80-95.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Table 1
Heuristic Data: Departure from Linearity

I.D.	Anxiety	Test Score
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7	7	7
8	8	8
9	9	9
10	10	10
11	11	9
12	12	8
13	13	7
14	14	6
15	15	5
16	16	4
17	17	3
18	18	2
19	19	1

Table 2
Descriptive Statistics and r for Table 1 Data

Variable	Mean	Variance	Skewness	r with X
Anxiety(X)	10.00	31.62	.000	-----
Test score(Y)	5.26	7.98	.031	0.000

Table 3
Heuristic data

I.D.	SATmath (X)	Mathtest (Y)	Yskewed	Ytimes5
1	340	58	58	290
2	390	63	59	315
3	480	72	60	360
4	490	73	61	365
5	510	75	62	375
6	520	76	63	380
7	520	76	63	380
8	530	77	65	385
9	530	77	65	385
10	550	79	67	395
11	550	79	67	395
12	570	81	69	405
13	570	81	69	405
14	580	82	72	410
15	580	82	72	410
16	590	83	78	415
17	610	85	80	425
18	620	86	90	430
19	710	95	95	475
20	760	100	100	500

Table 4
Descriptive Statistics and r for Table 3 Data

Variable	Mean	Variance	Skewness	r with X
SATmath (X)	550.00	8673.68	.0000	-----
Mathtest(Y)	79.00	86.74	.0000	1.0000
Yskewed	70.75	145.46	1.3210	0.8978
Ytimes5	395.00	2168.42	.0000	1.0000

Note. Although the Pearson r between X and Yskewed is +0.8978, because of difference in distribution shapes for the two variables, the Spearman's rho between the variables is 1.0000. This is because rho assumes no interval scale underlying the variables, and only evaluates whether the two variables order the cases in exactly the same way.

Table 5
Heuristic Data: Restricted Range Effects on r

I.D.	SATmath(X)	Mathscore(Y)	Xrestricted	Yrestricted
1		340	63	
2		340	65	
3		350	71	
4		470	74	
5		480	79	
6		490	80	
7		500	80	
8		510	84	
9		520	85	
10		570	85	
11	570	85	570	85
12	580	87	580	87
13	590	89	590	89
14	590	91	590	91
15	620	92	620	92
16	620	93	620	93
17	630	94	630	94
18	630	95	630	95
19	710	96	710	96
20	760	100	760	100

Table 6
Descriptive Statistics and r for Table 5 data

Variable	Mean	Variance	Skewness	r with X
SATmath (X)	543.50	12866.05	-.313	-----
Mathtest(Y)	84.40	104.04	-.664	.9722
Xrestricted	630.00	3644.44	1.413	-----
Yrestricted	92.20	19.73	.015	.9032

Figure 1
Scattergram Plot of Anxiety with Test Score

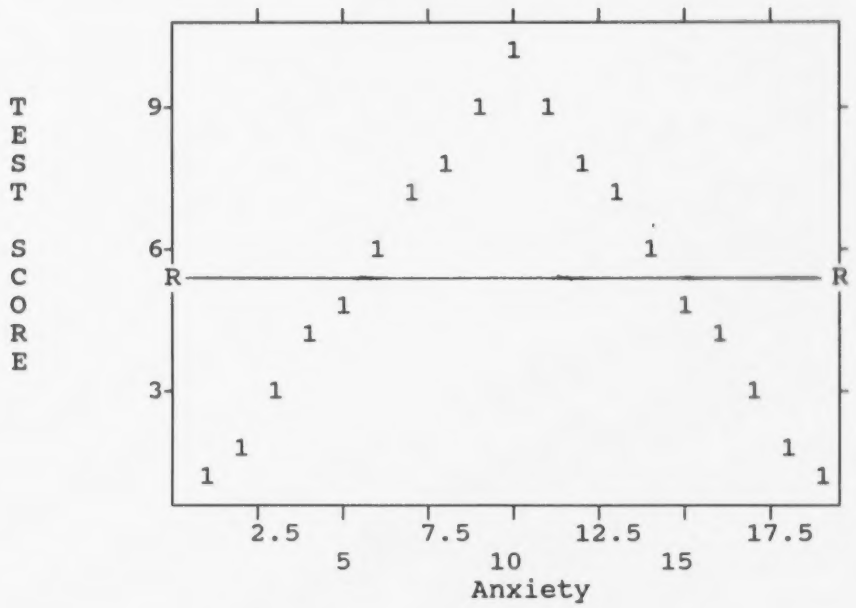


Figure 2
Distribution Shapes Affect r

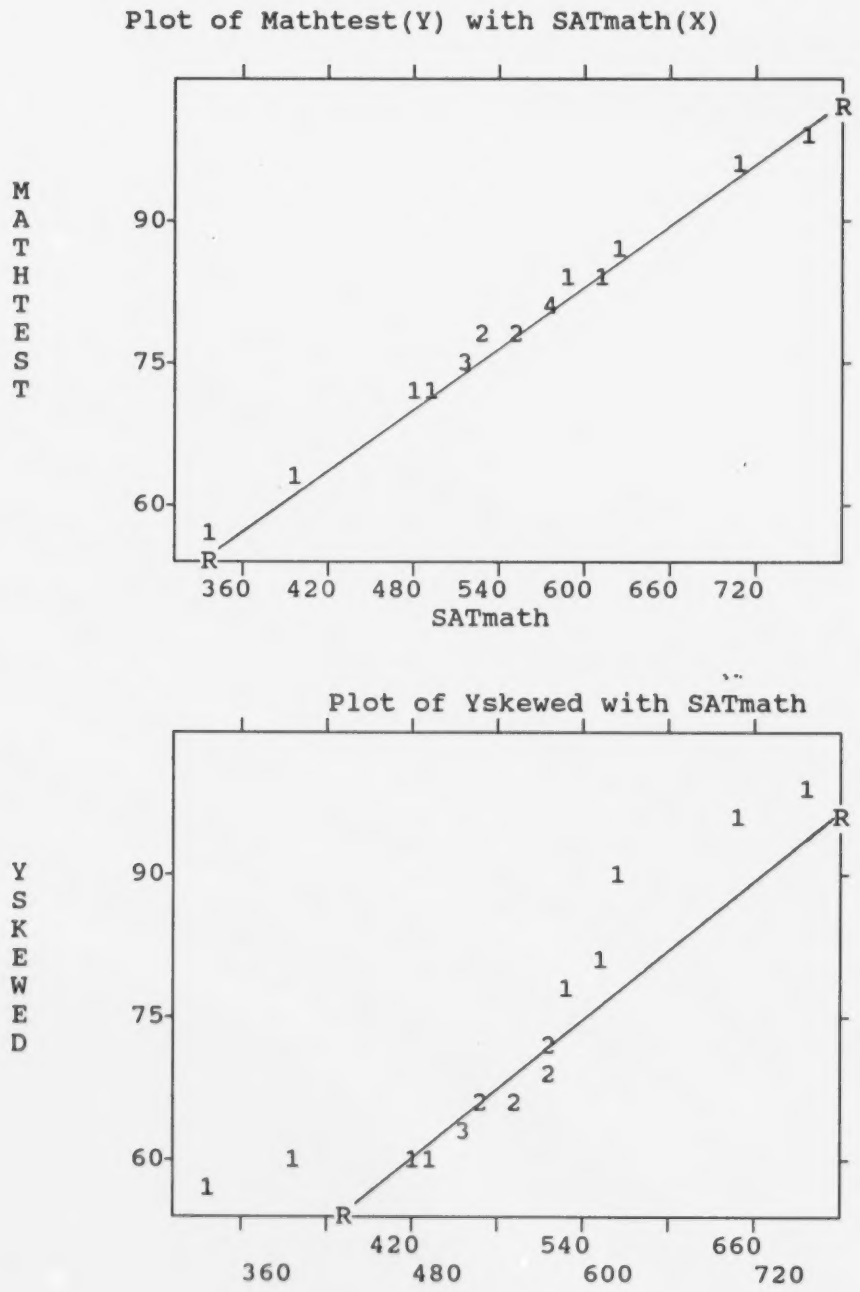


Figure 3
Artificial Variance Changes Do Not Affect r

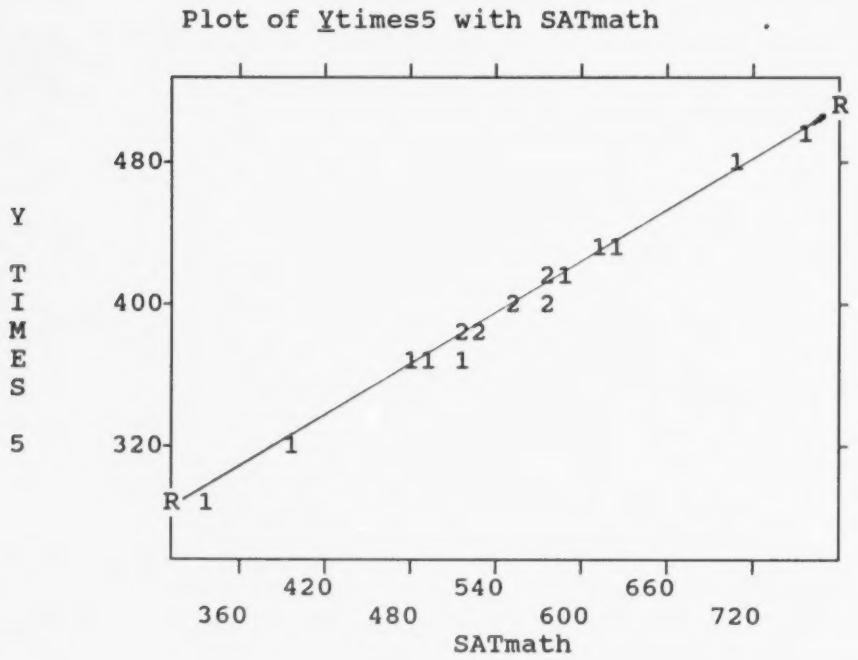
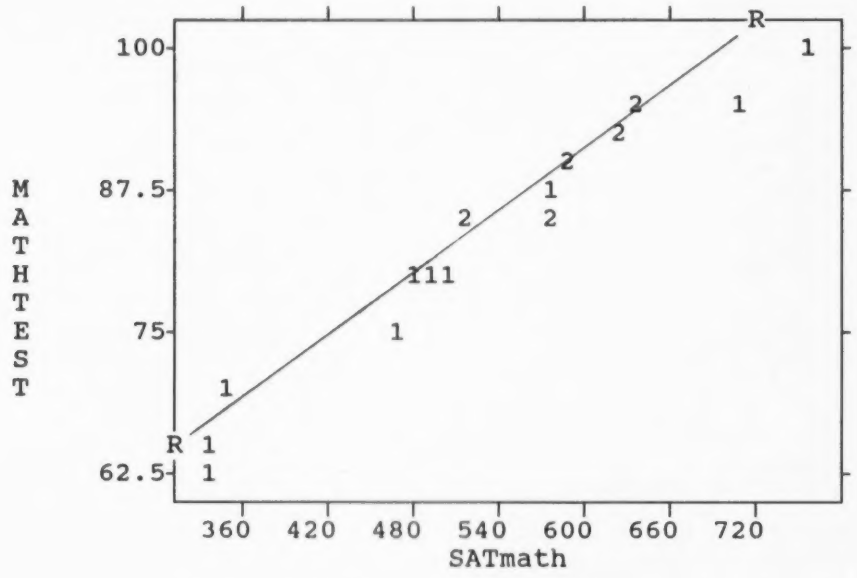


Figure 4
True Restricted Ranges Do Affect r

Plot of Mathtest with SATmath using all Table 5 data



Plot of Mathtest with SATmath